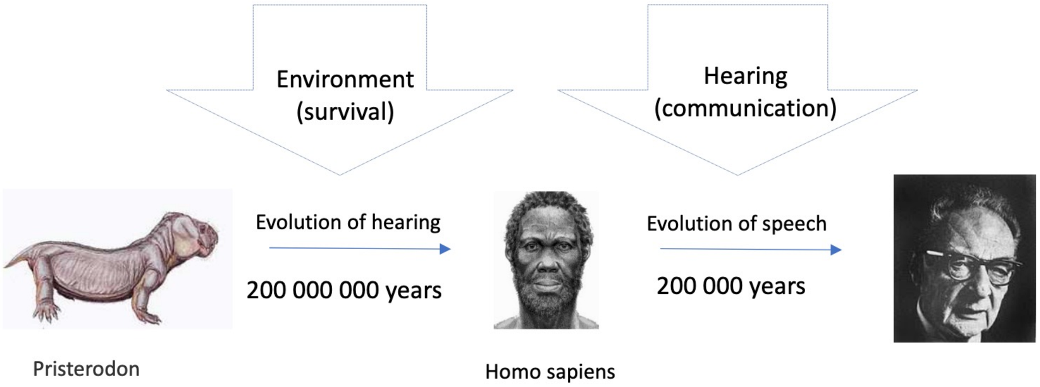
Speech basics

Instructor: Laureano Moro-Velazquez

Most slides from Hynek Hermansky



Environment (survival)

Hearing (communication)

Evolution of hearing

200 000 000 years

Evolution of speech

200 000 years

Pristerodon

Homo sapiens

We hear to survive

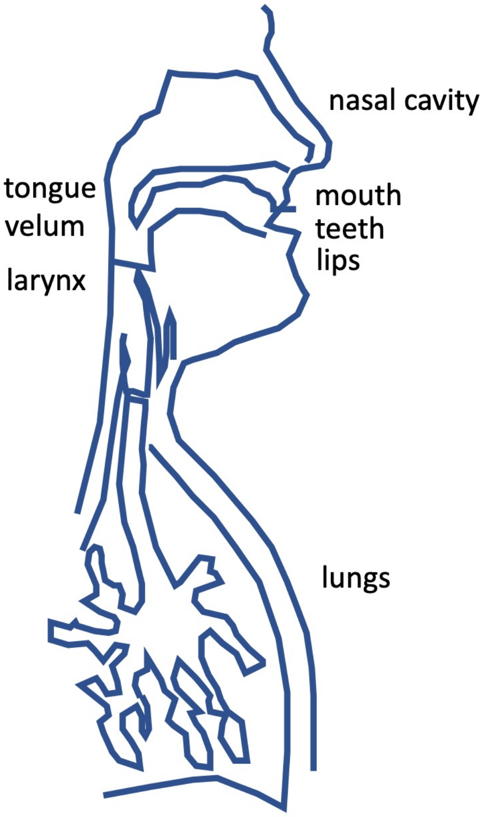
> …. sensory neurons are adapted to the statistical properties of the signals to which they are exposed.
>
> Simoncelli and Olshausen

We speak to hear

> **We speak in order to be heard** and need to be heard in order to be understood.
>
> Jakobson and Waugh p.95

Human speech evolved to fit properties of human hearing

Speech generation



breathing

eating

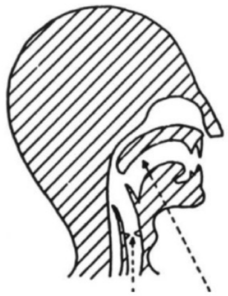
biting

speaking?

Speech – sequence of short stationary (10-20 ms) signal segments

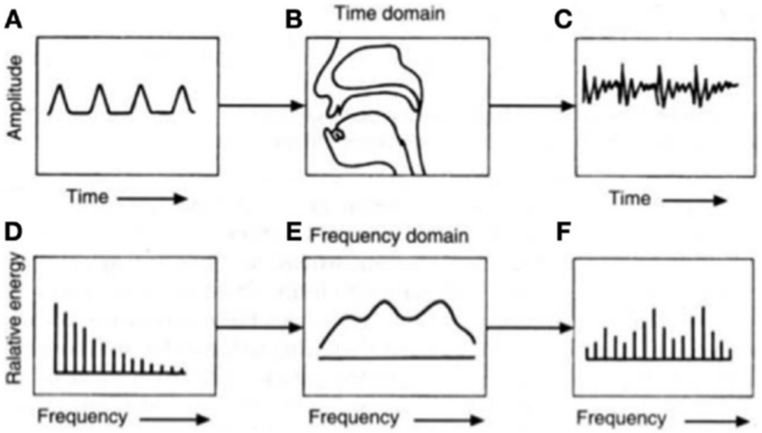
Linear model of speech production (Chiba and Kajiyama 194, Fant 1960, ..)

source → filter → filtered source signal



speech signal

source filter

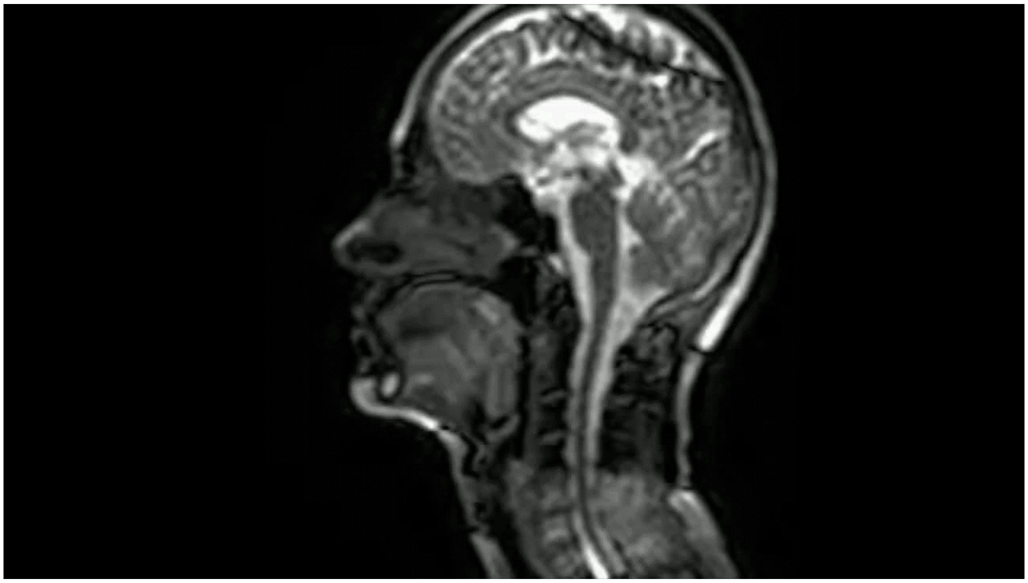


vocal source vocal tract speech signal

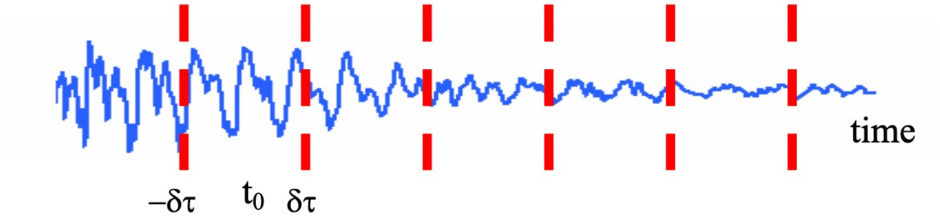
The source

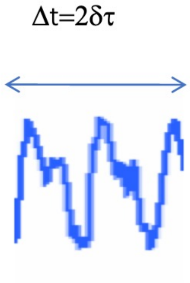
[redacted]

The vocal tract (the filter)



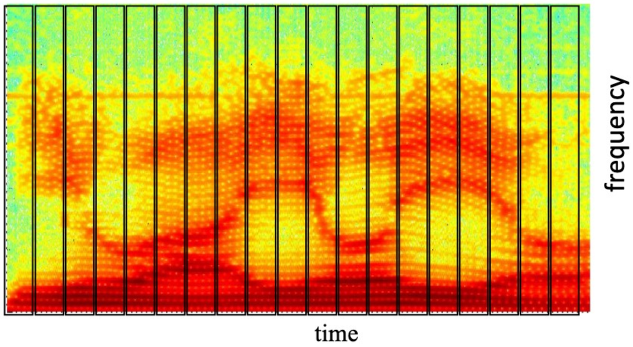
Video of rt-MRI of vocal tract during speech. (Freitas, A. C. et al, 2016) non-stationary speech signal s(t)

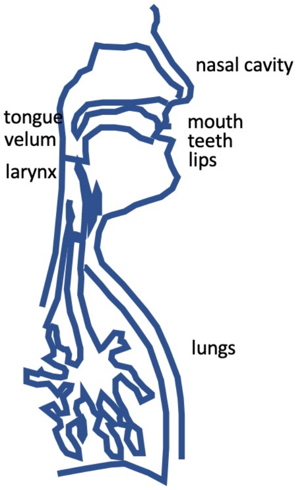




→ $\mathbf{S(\omega)}= \mathcal{F}\{s(t_0 \pm \delta\tau)\}$

time-frequency representation of the signal
(spectrogram)

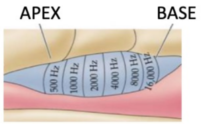




Redundant spread of information

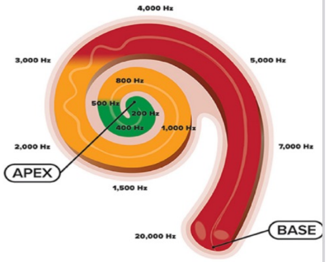
- every change of the tract shape shows at all frequencies of speech spectrum
- tract shape changes do not happen very fast

brain



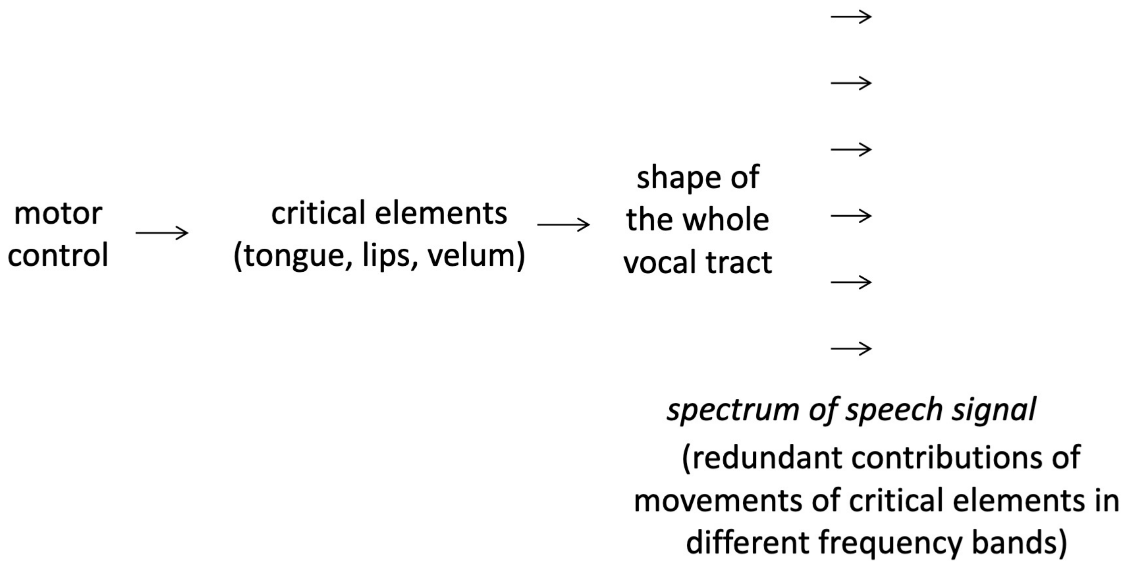
↑
Medial geniculate body
↑
Inferior colliculus
↑
Superior olive
↑
Cochlear nucleus
↑
Auditory nerve
↑

ear



- frequency selective (about 20 bands)
- sluggish (tenths of seconds)

INFORMATION ABOUT TRACT SHAPES DISTRIBUTED IN FREQUENCY



motor control → critical elements (tongue, lips, velum) → shape of the whole vocal tract →

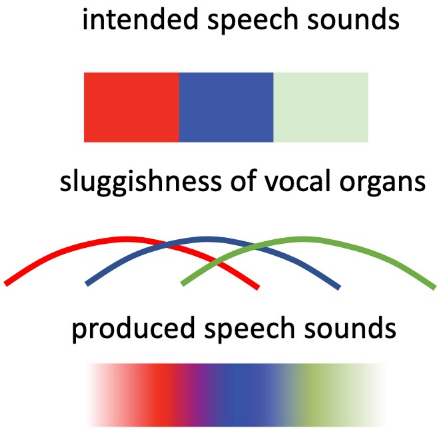
spectrum of speech signal

(redundant contributions of movements of critical elements in different frequency bands)

INFORMATION ABOUT TRACT SHAPES DISTRIBUTED IN TIME

from Sri Narajanan

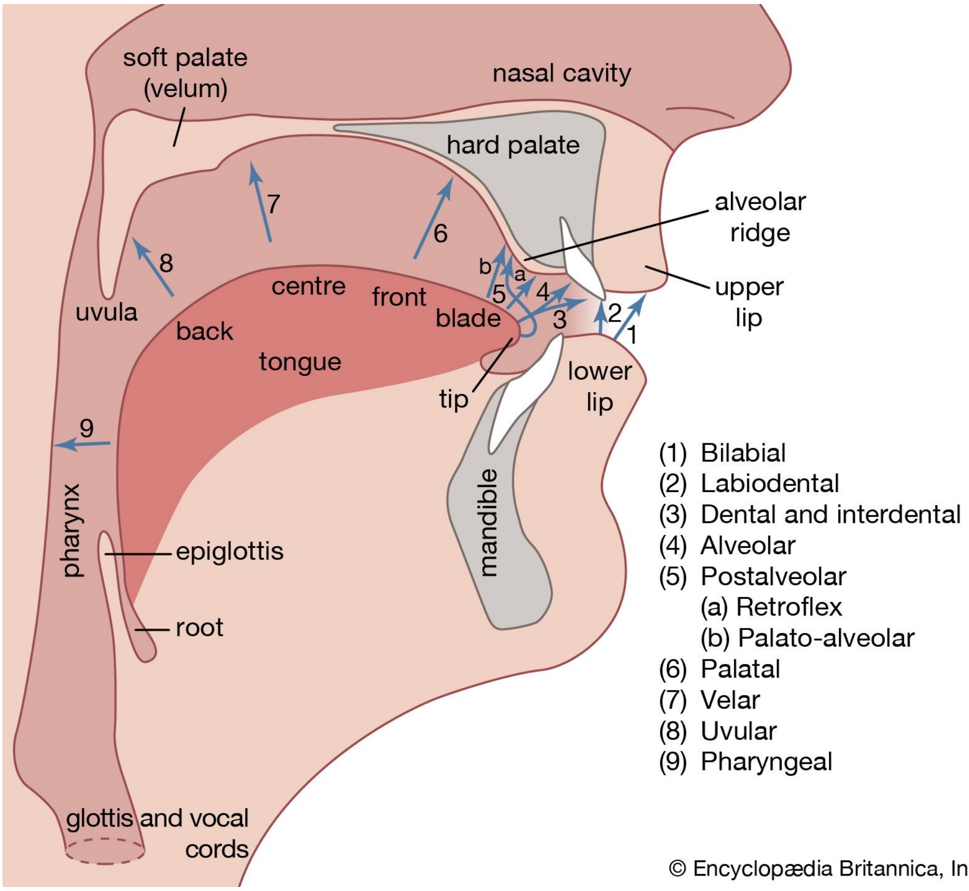
movements of vocal organs are
rather sluggish



intended speech sounds

sluggishness of vocal organs

produced speech sounds



© Encyclopædia Britannica, Inc.

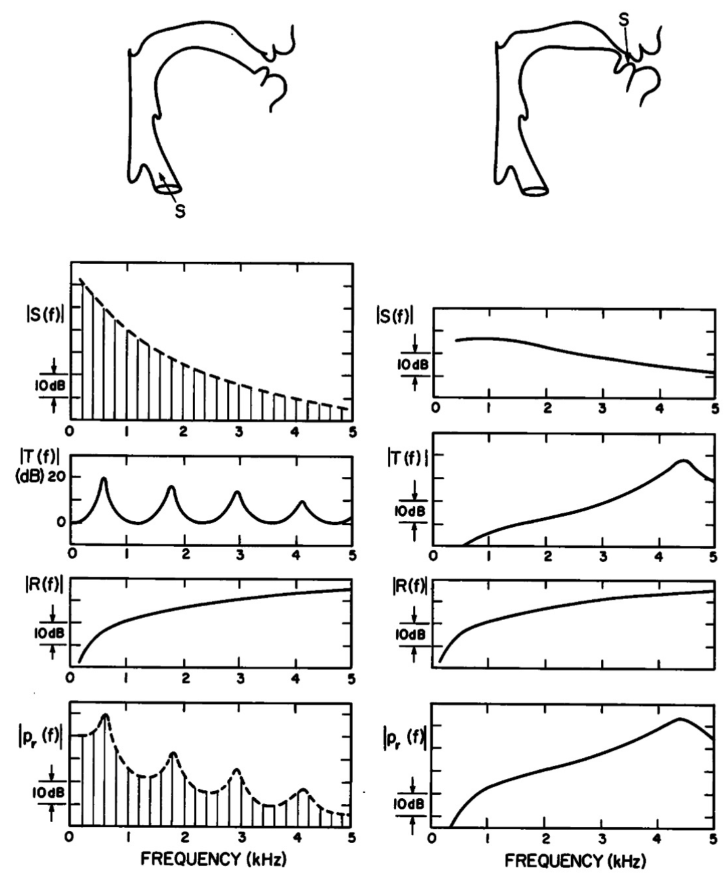
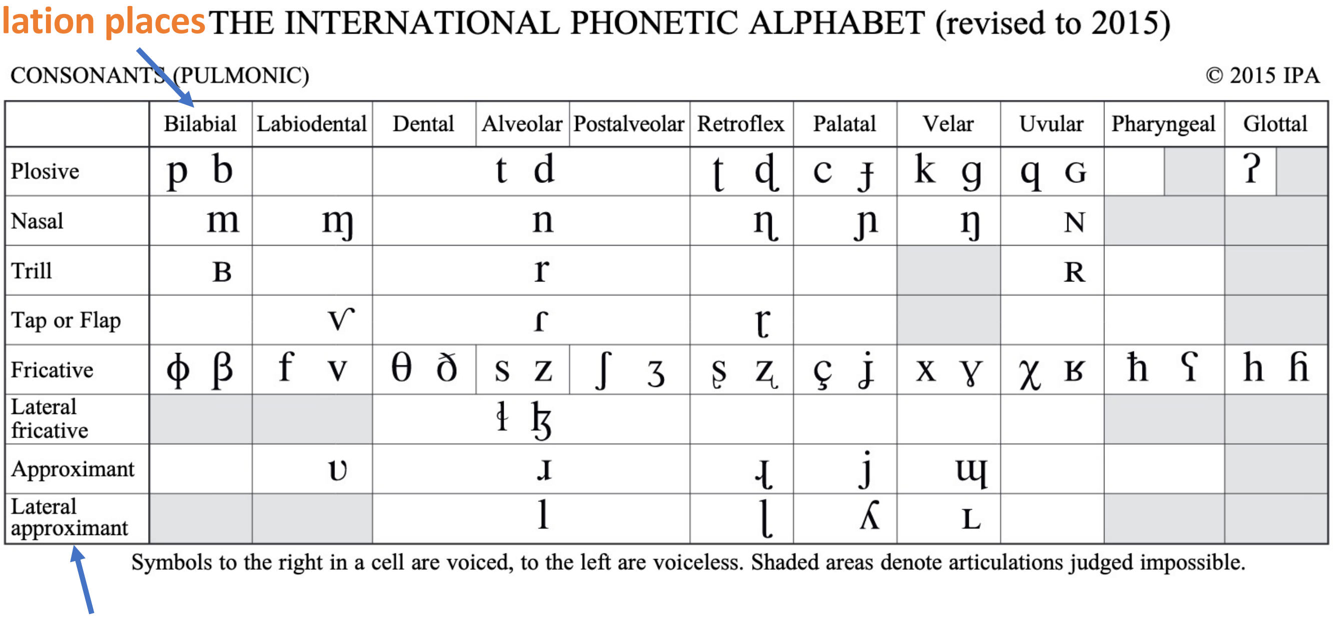


Figure 3.1 Sketches indicating components of the output spectrum $|p_r(f)|$ for a vowel and a fricative consonant. The output spectrum is the product of a source spectrum $S(f)$, a transfer function $T(f)$, and a radiation characteristic $R(f)$. The source spectra are similar to those derived in figures 2.10 and 2.33 in chapter 2. For the periodic source, $S(f)$ represents the amplitudes of spectral components; for the noise source, $S(f)$ is amplitude in a specified bandwidth. See text.

Articulation places



THE INTERNATIONAL PHONETIC ALPHABET (revised to 2015)

CONSONANTS (PULMONIC)

© 2015 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

Manner classes

Some phonetics

- **Articulation places:** points or areas in the vocal tract where there is a constriction (with or without contact) which has the most relevance in the generated sound.

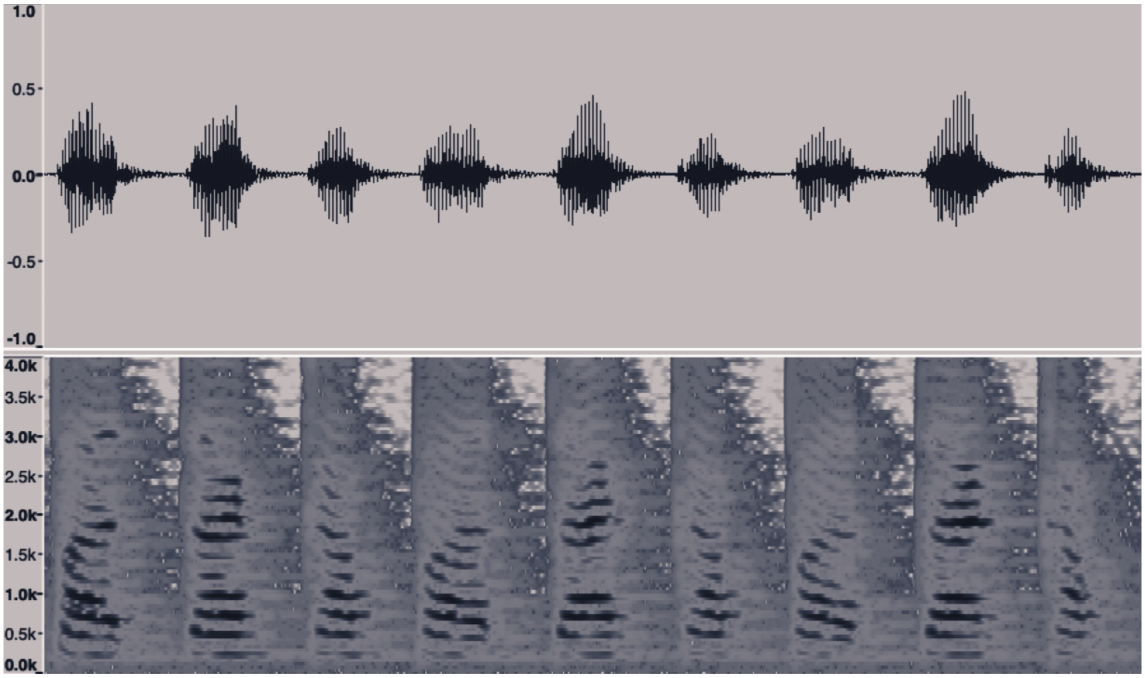
- **Manner classes:** ways in which we articulate that produce consonant sounds.

Some phonetics: manner classes

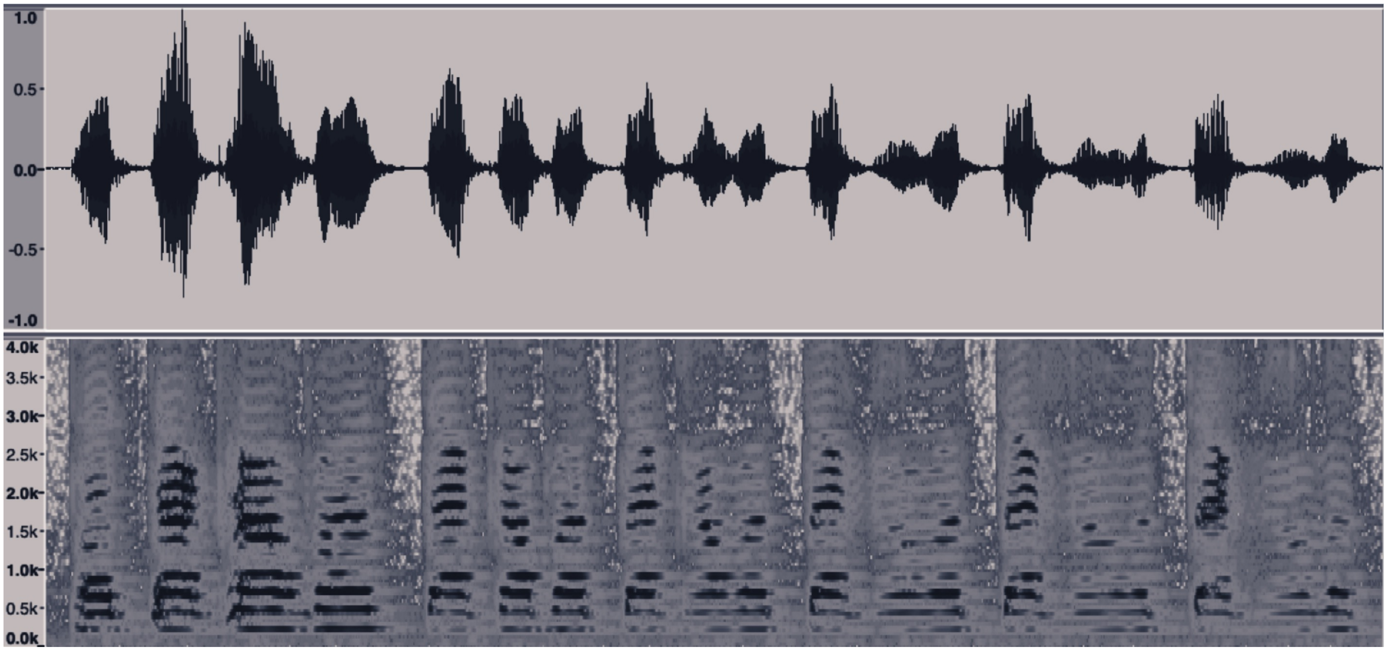
- Plosive: There is a constriction in the vocal tract and the airflow is interrupted for a short period of time (stop). Then, there is usually a release of the air that generates a sound.

- Fricative: The constriction of the vocal tract is very narrow but does not interrupt the airflow. This generates certain turbulences that lead to fricative sounds

Example of plosives + vowels



Combination of plosives, fricatives, and vowels

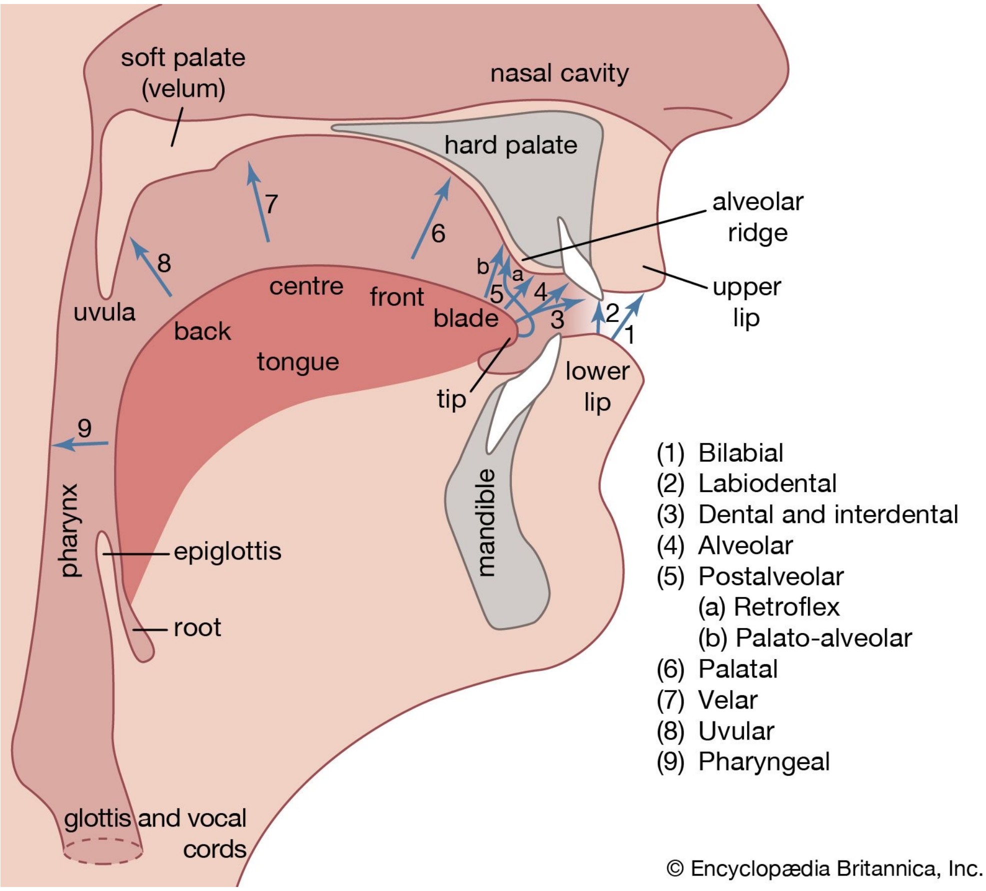


Some phonetics: manner classes

- Nasal: The soft palate is lowered, and the airflow goes through the nasal cavity. Nasal sounds are usually voiced (the source is on).

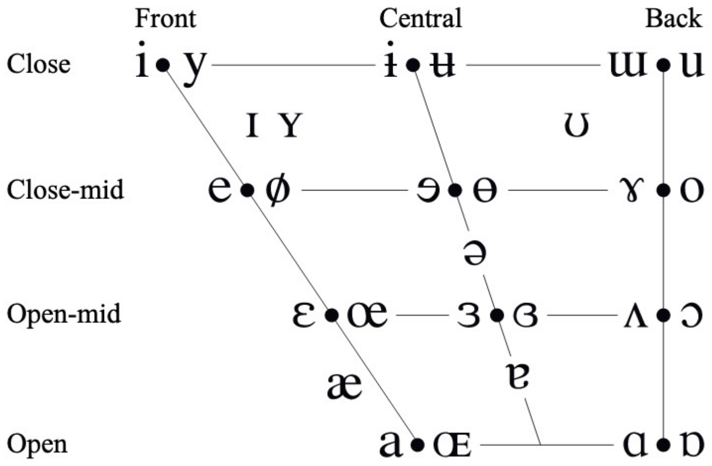
- Trill: The articulator (tongue, lips…) vibrate against other parts of the mouth with multiple flaps while the airstream is flowing (like in the word Ramon, in Spanish).

- Flap: Is similar to trill, but in this case there is only one short flap, like /r/ in the word radar in English).



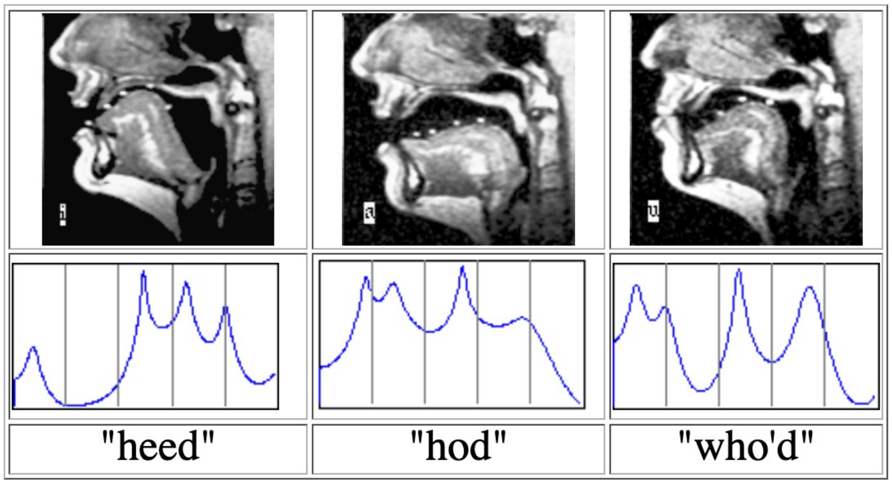
© Encyclopædia Britannica, Inc.

VOWELS



Where symbols appear in pairs, the one
to the right represents a rounded vowel.

Formants



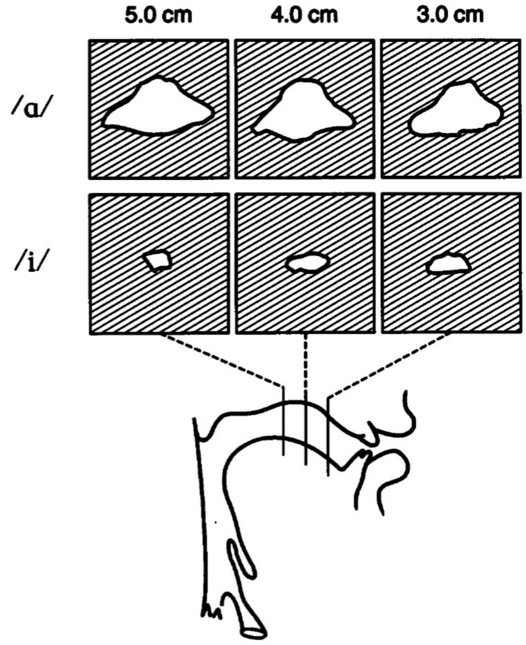
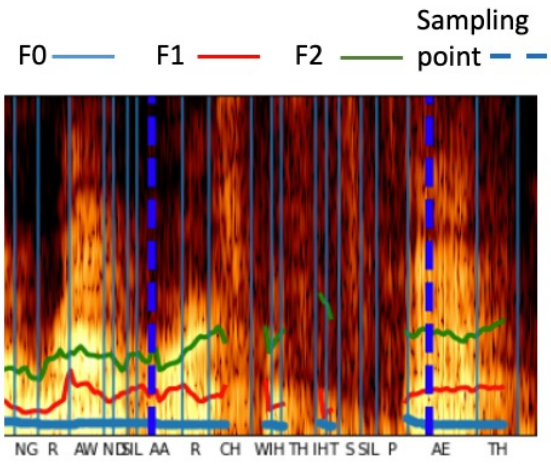


Figure 1.22 Cross-sectional shape of the airway at three different positions in the oral region when the vocal tract is configured to produce the vowels /ɑ/ and /i/. The distance of the section from the lips is marked on each panel. The approximate locations of the sections are shown in the midsagittal tracing at the bottom. (The midsagittal tracing is not taken from the same vocal tract as the coronal sections.) The coronal sections were obtained from a magnetic resonance imaging technique. (Data from Baer et al., 1991.)

Vowel Space Area



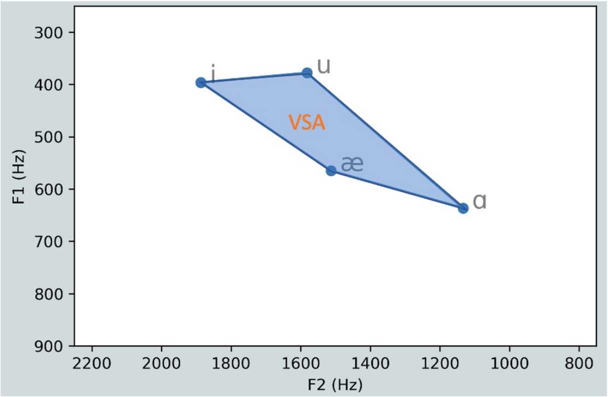
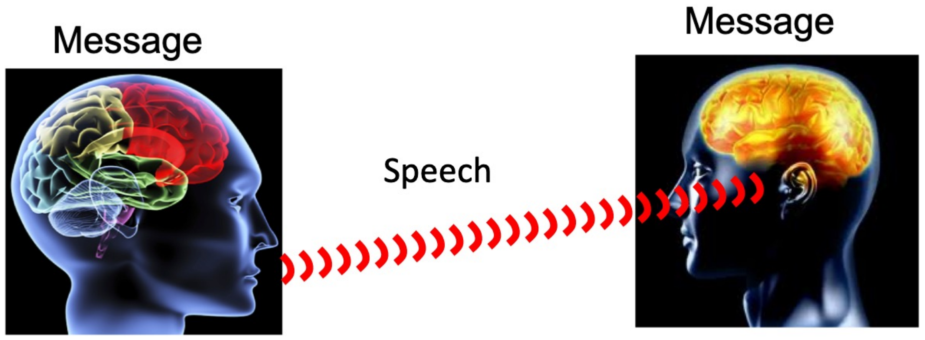


Fig 1. Quadratic Vowel Space Area in American English.

The message

Human Speech



Message

Speech

Message

Messages

- Only a limited number of speech sounds can be produced and distinguished
- Many things need to be said

Composionality: meaning of a complex expression is determined by the meanings of its constituent expressions and the rules used to combine them (*Wikipedia)*

Create words as ordered sequences of speech sounds (phonemes).
file /fīl/
life /līf/

Create phrases as ordered sequences of words.
Tom chased horse.
Horse chased Tom.

Prior probabilities of different letters in English alphabet

Letter	**Relative frequency**	**Letter**	**Relative frequency**
e	12.702%	**m**	2.406%
t	9.056%	**w**	2.360%
a	8.167%	**f**	2.228%
o	7.507%	**g**	2.015%
i	6.966%	**y**	1.974%
n	6.749%	**p**	1.929%
s	6.327%	**b**	1.492%
h	6.094%	**v**	0.978%
r	5.987%	**k**	0.772%
d	4.253%	**j**	0.153%
l	4.025%	**x**	0.150%
c	2.782%	**q**	0.095%
u	2.758%	**z**	0.074%

Samuel Morse
(self-portrait)

Morse code

e - single dot
z - dot and three dashes

In 1939, Ernest Vincent Wright published a 267-page novel, Gadsby, in which **no use is made of the letter E**. Here is a paragraph from the novel:

Upon this basis I am going to show you how a bunch of bright young folks did find a champion; a man with boys and girls of his own; a man of so dominating and happy individuality that Youth is drawn to him as is a fly to a sugar bowl. It is a story about a small town. It is not a gossipy yarn; nor is it a dry, monotonous account, full of such customary "fill-ins" as "romantic moonlight casting murky shadows down a long, winding country road." Nor will it say anything about tinklings lulling distant folds; robins carolling at twilight, nor any "warm glow of lamplight" from a cabin window. No. It is an account of up-and-doing activity; a vivid portrayal of Youth as it is today; and a practical discarding of that worn- out notion that "a child don't know anything."

How “efficient” is a given code?

Entropy

$$H(s) = -\sum_{i=1}^{n} p_i \cdot \log(p_i)$$

26 letters and space of English alphabet

$$H(s) = -\sum_{i=1}^{27} 1/27 \cdot \log(1/27)$$

$$= -\log(1/27) = 4.74 \text{ bit}$$

all letters are equally probable (zero order)

H(s)= 4.74 bit

Respecting relative frequencies of letters
(first order)
H(s)= 4.28 bit

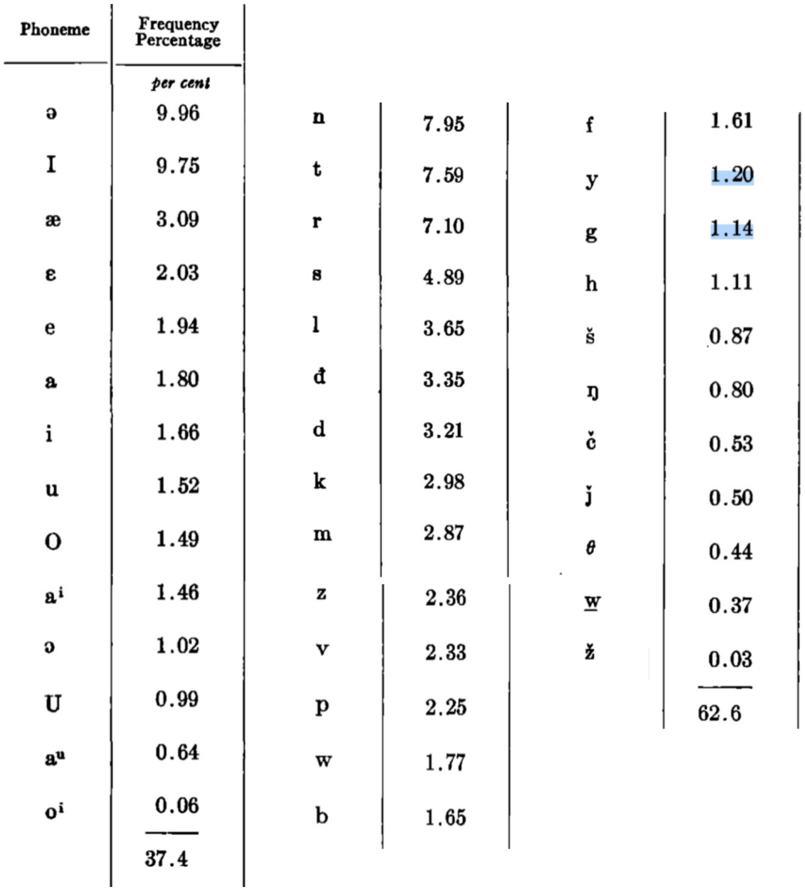
Respecting relative frequencies of combinations of three letters (third order)
H(s)= 2.77 bit

Letters in real text
(estimate)
H(s) ~ 0.6-1.3 bit

Shannon
Prediction and Entropy of Printed English
BSTJ 1951

The Relative Frequency of Phonemes in General-American English

Hayden 1950



Phoneme	Frequency Percentage
	per cent
ə	9.96
I	9.75
æ	3.09
ɛ	2.03
e	1.94
a	1.80
i	1.66
u	1.52
O	1.49
a^{i}	1.46
ɔ	1.02
U	0.99
a^{u}	0.64
o^{i}	0.06
	37.4
n	7.95
t	7.59
r	7.10
s	4.89
l	3.65
đ	3.35
d	3.21
k	2.98
m	2.87
z	2.36
v	2.33
p	2.25
w	1.77
b	1.65
f	1.61
y	1.20
g	1.14
h	1.11
š	0.87
ŋ	0.80
č	0.53
ǰ	0.50
θ	0.44
w̲	0.37
ž	0.03
	62.6

Phonemes

Perceptually distinct speech elements that could distinguish one words from another

Graphemes

Letters and combinations of letters representing speech sounds (phonemes)

Rotokas language – East of New Guinea, 11 phonemes,
12 symbols, 1 symbol per sound

Taa language – Botswana (Africa), ~ 200 phonemes ,
20-22 symbols, up to 6 symbols per sound

English
~45 phonemes, 27 symbols,
~ 250 graphemes, up to 5 symbols per sound

40 speech sounds (phonemes) in American English
24 consonants
19 vowels and diphtongs

vowels – mouth open
consonants - mouth not so open

typical syllable

cvc
onset – nucleus – coda
cv
onset – nucleus

/l/,/r/,/w/,/y/ - semivowels
produced with open mouth
can stand as nucleus in syllable

Phones, phonemes and allophones

- **Phone:** we usually call phone to a specific segment that contains a distinct sound, but it does not have to be critical to the meaning of a word. A phone can be a phoneme or part of it.

- **Allophones:** different realizations of a phone, depending on the dialect or other domain changes. These do not change the word meaning when they change.

Phones, phonemes and allophones

- If in a word you change a phoneme, you will change the meaning of a word. If you change a phone, you might not change the meaning of that word.

- The phoneme is the mental realization, the phone is the sound representation of a phone.

Phonotactics

- Phonotactics: the study of the rules governing the possible phoneme sequences in a language. (Oxford Dictionary).
- Phonotactic rules define permissible phoneme sequence (syllable) structure in a language.
- It deals with restrictions in a language on the allowed (or expected) combinations of phonemes.
- For instance, you will not find this combination of sounds in English: /nnnnisffg/.

Words

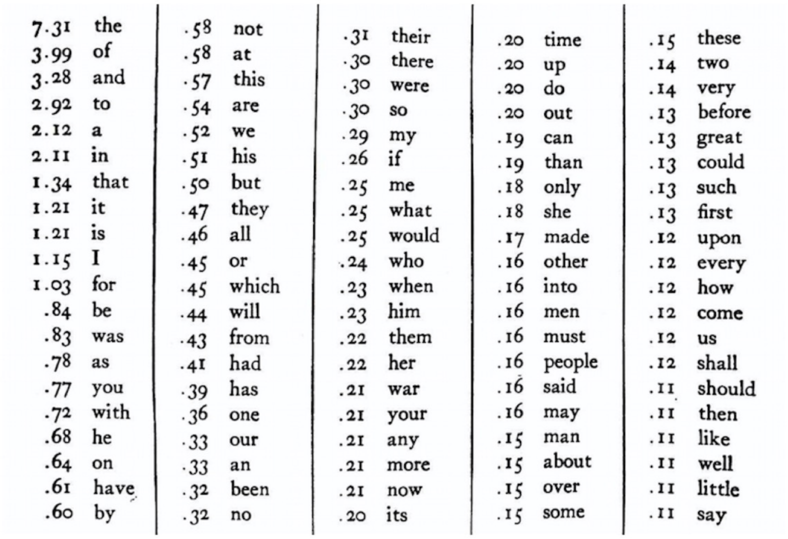
- ordered combinations of speech sounds
- represent objects, ideas, actions, relationships, qualities, e.t.c., **as agreed on by a particular society (language)**
- new words constantly invented and old words changing their meanings
- learned using interventions and rewards from other human beings
- particular word meanings often depend on context

Word sequences (sentences, phrases,..)

- Words organized into larger units (sentences, phrases,..) using rules of the language (syntax, grammar)

- Order also carries information
    - John beats Frank. Frank beats John.
    - I went home and had a dinner. I had a dinner and went home.

Relative frequencies of words in written English [%]



7.31	the	.58	not	.31	their	.20	time	.15	these
3.99	of	.58	at	.30	there	.20	up	.14	two
3.28	and	.57	this	.30	were	.20	do	.14	very
2.92	to	.54	are	.30	so	.20	out	.13	before
2.12	a	.52	we	.29	my	.19	can	.13	great
2.11	in	.51	his	.26	if	.19	than	.13	could
1.34	that	.50	but	.25	me	.18	only	.13	such
1.21	it	.47	they	.25	what	.18	she	.13	first
1.21	is	.46	all	.25	would	.17	made	.12	upon
1.15	I	.45	or	.24	who	.16	other	.12	every
1.03	for	.45	which	.23	when	.16	into	.12	how
.84	be	.44	will	.23	him	.16	men	.12	come
.83	was	.43	from	.22	them	.16	must	.12	us
.78	as	.41	had	.22	her	.16	people	.12	shall
.77	you	.39	has	.21	war	.16	said	.11	should
.72	with	.36	one	.21	your	.16	may	.11	then
.68	he	.33	our	.21	any	.15	man	.11	like
.64	on	.33	an	.21	more	.15	about	.11	well
.61	have	.32	been	.21	now	.15	over	.11	little
.60	by	.32	no	.20	its	.15	some	.11	say

In spoken language most frequency word is pronoun “I”

Telephone conversations 5%

Schizophrenics 8.4%

Predictability and unpredictability

- 100 % predictable message has no information value
  - When knowing exactly what will be said, no need to listen

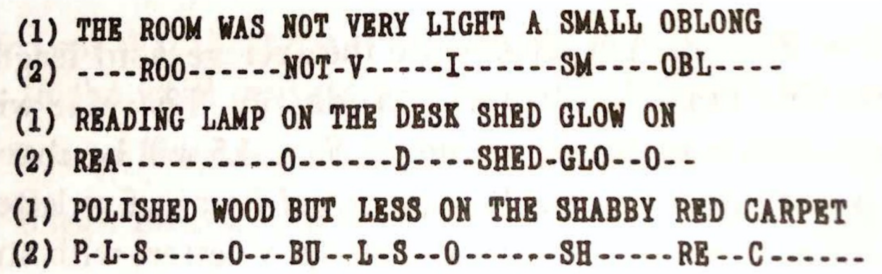
- Speech is to large extent predictable since is follows rules
  - Grammar, use of words, word order, …

- The predictability allows for easier communication

To communicate effectively, the right balance between predictability and unpredictability need to be maintained.

How predictable is language? - Claude Shannon

1. Think about the English sentence
2. Ask people to think about the first letter in the sentence
3. When correct, tell them, mark it by "-" and ask for the second letter
4. When incorrect, tell them the correct one and ask for the second letter
5. Go on until the end of the sentence



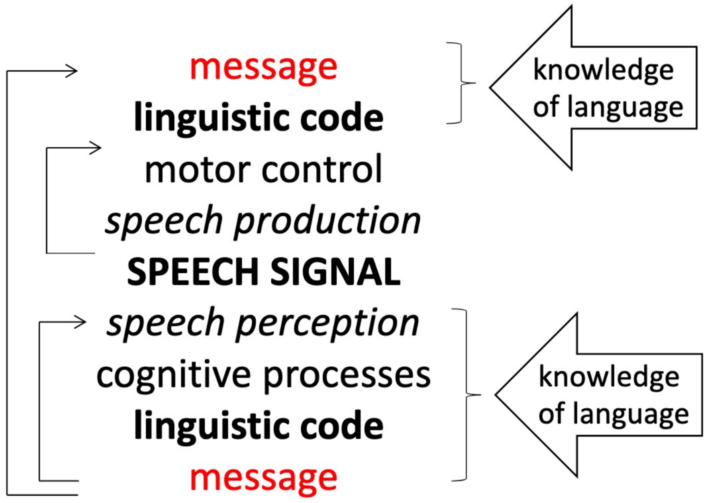
```
(1) THE ROOM WAS NOT VERY LIGHT A SMALL OBLONG
(2) ----ROO------NOT-V-----I------SM----OBL----
(1) READING LAMP ON THE DESK SHED GLOW ON
(2) REA----------O------D----SHED-GLO--O--
(1) POLISHED WOOD BUT LESS ON THE SHABBY RED CARPET
(2) P-L-S-----O---BU--L-S--O------SH-----RE--C------
```

69% of letters guessed correctly

Both line (1) and (2) contain the same information

- The line (1) can be guessed from the info in the line (2) – by the identical twin ☺

Human Speech



message

linguistic code

motor control

speech production

SPEECH SIGNAL

speech perception

cognitive processes

linguistic code

message

knowledge of language

knowledge of language

27 symbols of English alphabet
~15 symbols/s, each symbol ~ 3 bits/symbol
> 50 b/s

standard PCM coding
8 kHz sampling, 11 bit accuracy
88 kb/s

> 50 b/s

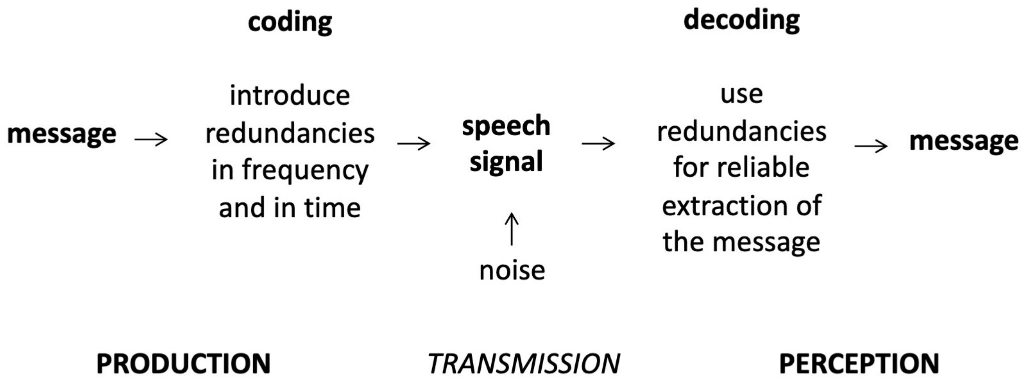
INFORMATION in speech signal: **message,** who is speaking, health, language, emotions, mood, social status, acoustic environment, etc,…

Why speech?

- Profit
  - searching large speech databases, transcription, voice control,…
- *voice will do to touch what touch did to keyboards*.
  - Mooly Eden, senior vice president Intel

- Important spin-offs
  - Digital signal processing
  - Sequence classification (Hidden Markov Models)
    - financial predictions
    - human DNA matching
    - action recognition
  - Image processing techniques

Spoken language is one of the most amazing accomplishments of human race.



coding

decoding

message → introduce redundancies in frequency and in time → **speech signal** → use redundancies for reliable extraction of the message → **message**

noise ↑

PRODUCTION

TRANSMISSION

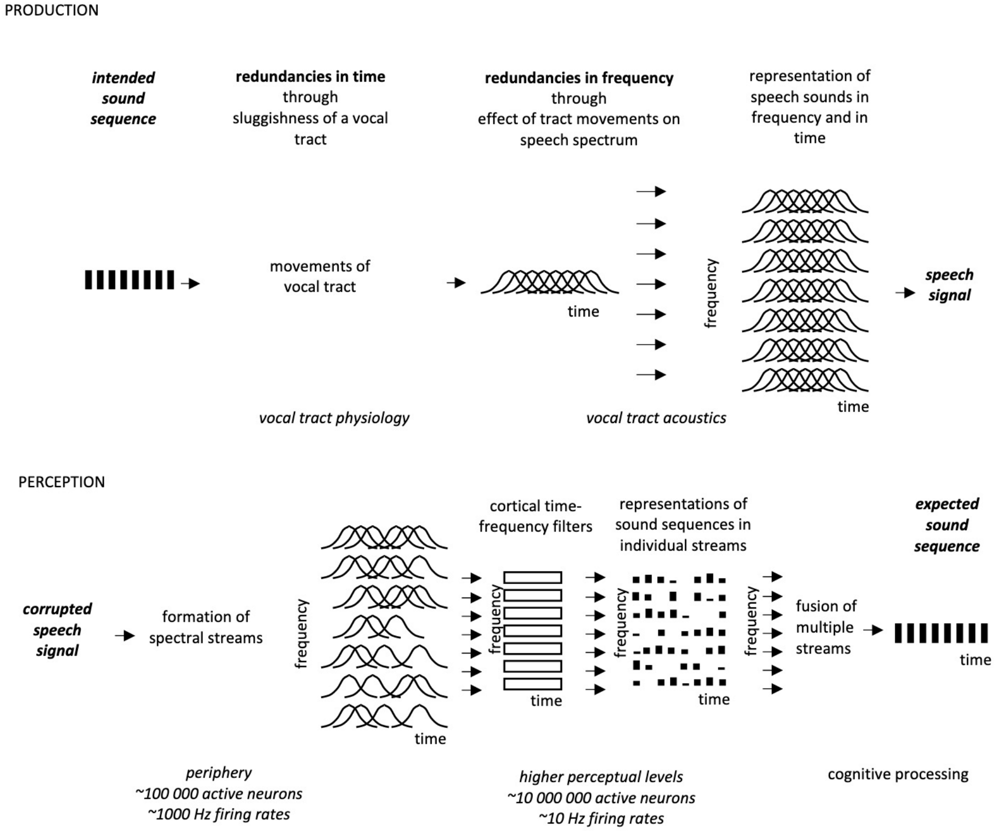
PERCEPTION

redundancy in frequency

production*:* tract acoustics distributes the information to all frequencies of the speech spectrum
perception: hearing selectivity allows for decoding the information in separate frequency bands

redundancy in time

production: tract sluggishness (coarticulation) distributes information about each speech sound in time
perception: temporal sluggishness of hearing collect the information distributed in time


PRODUCTION
intended sound sequence
redundancies in time through sluggishness of a vocal tract
redundancies in frequency through effect of tract movements on speech spectrum
representation of speech sounds in frequency and in time
movements of vocal tract
time
frequency
time
speech signal
vocal tract physiology
vocal tract acoustics
PERCEPTION
corrupted speech signal
formation of spectral streams
frequency
time
cortical time-frequency filters
frequency
time
representations of sound sequences in individual streams
frequency
time
frequency
fusion of multiple streams
expected sound sequence
time
periphery
~100 000 active neurons
~1000 Hz firing rates
higher perceptual levels
~10 000 000 active neurons
~10 Hz firing rates
cognitive processing

Human Language Technologies

A brief look

October 4
Hi, how can I help?

Are We There Yet ?

- Repetition, fillers, hesitations, interruptions, unfinished and non-grammatical sentences, new words, dialects, emotions, …
- Hands-free operation in noisy and reverberant environments,…

Alleviate need for large amounts of annotated training data

- Robustness to speech distortions, which do not seriously impact human speech communication
- Dealing with new unexpected lexical items
- Unsupervised learning/adaptation?

How to Get There ?

Fred Jelinek

Speech recognition
…a problem of maximum likelihood decoding

information and communication theory, machine learning, large data,….

Roman Jakobson

We speak, in order to be heard, in order to be understood

human communication, speech production, perception, neuroscience, cognitive science,..

Gordon Moore

The complexity for minimum component costs has increased at a rate of roughly a factor of two per year…

John Pierce

..devise a clear, simple, definitive experiments. So a science of speech can grow, certain step by certain step.

Signal processing,
information theory,
machine learning, …

&

neural information processing,
psychophysics, physiology,
cognitive science, phonetics and
linguistics, ...

Engineering and Life Sciences together !